Vocabularios para la descripción de conjuntos de datos en la Web

Vocabularies for describing datasets on the Web

Juan-Antonio Pastor-Sánchez

Pastor-Sánchez, Juan-Antonio (2017). "Vocabularios para la descripción de conjuntos de datos en la Web". *Anuario ThinkEPI*, v. 11, pp. 278-283.

https://doi.org/10.3145/thinkepi.2017.54

Publicado en *IweTel* el 13 de diciembre de 2016

Resumen: Las tecnologías de la web semántica han alcanzado un alto grado de desarrollo. Asimismo existen cientos de esquemas de metadatos y ontologías y miles de conjuntos de datos disponibles para su uso público. El siguiente paso es mejorar el acceso, reutilización y confianza en la calidad de los conjuntos de datos. En esta nota se muestran algunos vocabularios y ontologías para describir conjuntos de datos (DCAT) y su procedencia (PROV-O), especificar los mecanismos de acceso y uso de los datos (DUV) y definir indicadores de calidad de los mismos (DQV).

Palabras clave: Vocabularios; Metadatos; Ontologías; Web semántica; Conjuntos de datos; DCAT; PROV-O; DUV; DQV.

Abstract: The technologies of the semantic web have reached a high degree of development. There are hundreds of metadata schemes and ontologies and thousands of datasets available for public use. The next step is to improve access, reuse and trust on datasets. This article includes vocabularies for describing datasets (DCAT) and their provenance (PROV-O), specifying mechanisms for access and use of data (DUV), and defining quality indicators (DQV).

Keywords: Metadata; Vocabularies; Ontologies; Semantic web; Datasets; DCAT; PROV-O; DUV; DQV.

Introducción

El desarrollo y madurez de las tecnologías de la web semántica han ayudado a delimitar su despliegue en aplicaciones muy concretas y ha permitido implementar herramientas para publicar datos estructurados en la Web. La disponibilidad de dichos datos se realiza partiendo del principio de interoperabilidad semántica. Este es el objetivo de estándares como RDF, OWL o *Sparql* entre otros.

Son miles los conjuntos de datos que hacen uso de dichas tecnologías para su publicación y acceso. Desde este punto de vista, hay que considerar que la interoperabilidad no solamente reside en las tecnologías mencionadas anteriormente, sino también en la definición de vocabularios[1]. A este respecto resultan muy interesante iniciativas como *LOV*[2] que permite a los programadores buscar vocabularios ya existentes antes de proceder a definir otros nuevos que resuelvan necesidades específicas de descripción y representación de recursos. Es decir, la reutilización no sólo se centra en los conjuntos de datos, sino también en los vocabularios.

El grupo de trabajo *Data on the Web* del *W3C* incide en este punto en su documento sobre buenas prácticas (**Farias-Lóscio**; **Burle**; **Calegari**, 2016): la reutilización de vocabularios incrementa la interoperabilidad de los datos. En consecuencia, un conjunto de datos debe incluir información adicional a los propios datos que contiene siendo preciso añadir metadatos sobre

el propio conjunto de datos. Dichos metadatos deben abarcar aspectos descriptivos, estructurales, de interconexión con otros conjuntos de datos, el modo en el que pueden utilizarse, así como criterios e indicadores relacionados con la calidad del mismo.

Comprendiendo el concepto de conjunto de datos

Un conjunto de datos es algo más que un contenedor de sentencias RDF. Es una unidad autodescriptiva con datos estructurados disponibles en varios formatos para su descarga completa o acceso selectivo. El versionado de un conjunto de datos debe contemplar las actualizaciones a lo largo del tiempo. También es preciso introducir el concepto de distribución que facilita el uso de un conjunto de datos por parte de grupos de usuarios o consumidores de datos según sus necesidades. Es posible que en algunos casos se prefiera descargar el conjunto de datos completo en RDF/XML, utilizar un *Sparql Endpoint* en otros, e incluso habrá casos en los que se prefiera utilizar una API. Un conjunto de datos puede tener varias versiones a lo largo del tiempo y cada versión puede tener asociadas una o varias distribuciones.

Los metadatos que describen diversos aspectos de un conjunto de datos deben proporcionar información sobre el mismo, así como sobre sus versiones y diferentes distribuciones con el objetivo de facilitar la confianza en los datos y su reutilización. Adicionalmente a la descripción y la estructura del conjunto de datos dicha información debe incluir también información sobre posibles accesos a los datos, indicadores de calidad, procedencia y licencia de uso de los mismos.

Además todos los recursos (tanto el conjunto de datos en su totalidad como los elementos específicos) deben publicarse con IRIs[3] estables para que puedan ser referenciados. Los editores deben contemplar nuevas versiones del conjunto de datos y también tener en cuenta que el borrado sin más de los datos no es una buena práctica, siendo imprescindible indicar que el recurso ha sido borrado o archivado proporcionando información adecuada sobre ello[4]. Aquí es donde cobra especial importancia el concepto grafo RDF y el uso que se hace de las IRIs:

- es preciso aclarar que los recursos RDF no son los únicos elementos que pueden identificarse mediante una IRI. Un conjunto de datos en su totalidad también es identificado de esta forma;
- un conjunto de datos es en realidad una colección de grafos RDF.

Todos los grafos del conjunto de datos, excepto uno, tienen asociada una IRI. Estos grafos se denominan grafos con nombre (*named graphs*). El grafo que no tiene asociada una IRI se conoce como grafo por defecto (*default graph*) (**Bizer**, 2007, p. 62). Las diferentes sentencias RDF están asociadas a un determinado grafo con nombre para lo que se amplía el concepto de *RDF triples* (<sujeto><predicado><objeto>) con el de *RDF quads* (<sujeto><predicado><objeto><grafo>). Una sentencia que no esté asociada a un grafo con nombre implica que pertenece al grafo por defecto (**Carothers**, 2014).

La explicación anterior puede resultar un tanto técnica, pero es esencial para comprender como funciona el versionado de los conjuntos de datos. Aunque todavía no existe una forma normalizada, suele adoptarse el siguiente criterio: las sentencias RDF del grafo por defecto se corresponden con la versión más actualizada de los datos mientras que las sentencias correspondientes a versiones anteriores del conjunto de datos se vinculan a un grafo con nombre. Es necesario mantener versiones anteriores si se desea realizar una preservación digital planificada ya que esta práctica constituye una herramienta de auditoría y evaluación del conjunto de datos y un indicador de confianza.

"Mantener versiones anteriores de conjuntos de datos es una herramienta para su auditoría y evaluación, así como un indicador de confianza"

Describiendo los conjuntos de datos

Es preciso describir determinados aspectos del contenido y estructura del conjunto de datos. Existen dos enfoques al respecto:

- incluir, además de los propios datos, los oportunos metadatos autodescriptivos;
- gestión de un catálogo de conjuntos de datos con fines de control y búsqueda.

Para la primera tarea se suele utilizar VoID (*Vocabulary of Interlinked Datasets*), cuyo objetivo es describir no solamente el conjunto de datos en sí, sino hacer lo propio con los enlaces RDF que se establecen entre diferentes conjuntos de datos (**Alexander** *et al.*, 2011; **Cyganiak** *et al.*, 2011). Desde el punto de vista de RDF estos enlaces son aquellos cuyo sujeto y objeto están descritos en diferentes conjuntos de datos. Para describir las IRIs asociadas a un conjunto de datos, editor, fuente, fechas de creación, publicación o modificación, etc, se suelen utilizar los términos de metadatos de *Dublin Core*[5] o *FOAF*[6].

VoID ofrece una serie de clases y propiedades para especificar los vocabularios utilizados en el conjunto de datos, las particiones (subconjuntos

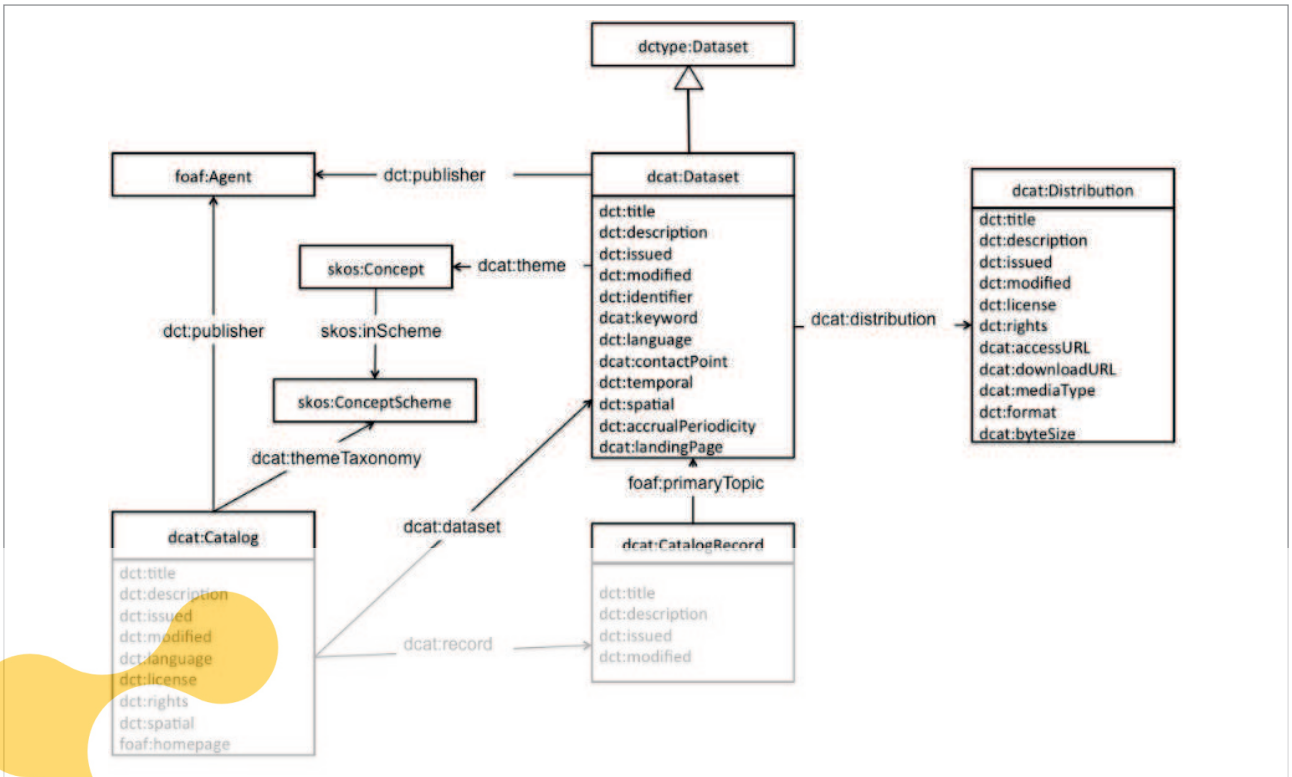

Figura 1. Modelo de datos de DCAT
https://www.w3.org/TR/vocab-dcat/dcat-model.jpg

de datos), las clases y propiedades utilizadas en el mismo e información estadística sobre su estructura. También permite identificar mecanismos de acceso mediante descarga, *Sparql Endpoint*, restricciones que proporcionan conjuntos de acceso, ejemplos de recursos, etc. Finalmente, *VoID* también permite describir los enlaces entre conjuntos de datos identificando los conjuntos de datos conectados y el tipo y número de enlaces existentes entre ellos.

Para la construcción de catálogos de conjuntos de datos se utiliza DCAT (*Data Catalog Vocabulary*). Se trata de un vocabulario para describir conjuntos de datos en catálogos de forma interoperable de manera que las aplicaciones puedan consumir datos descriptivos de los mismos de varios catálogos de forma simultánea (**Maali**; **Erickson**, 2014). Con DCAT es posible definir catálogos de registros con descripciones de conjuntos de datos. Otros aspectos de interés son la referenciación de los editores de los conjuntos de datos en el propio catálogo y el uso de vocabularios *SKOS* para clasificar los conjuntos de datos según su tema. La figura 1, sobre el modelo de datos de DCAT ilustra perfectamente la estructura de este vocabulario.

Procedencia

Los conjuntos de datos son objetos de información generados por un editor. A veces el conjunto de datos puede ser el reflejo de determinadas actividades. En este contexto, se entiende por Procedencia aquella información sobre organizaciones, actividades y personas que producen un conjunto de datos. Esta información puede ser utilizada como un medio para evaluar su calidad, fiabilidad y confianza. La ontología PROV (PROV-O) permite representar esta información (**Lebo**; **Sahoo**, 2013).

PROV-O identifica los conjuntos de datos como entidades que son generadas por actividades (que se llevan a cabo en un período de tiempo) y que a su vez se asocian a un agente. Lo interesante de la ontología PROV-O es que las diferentes clases y propiedades se organizan en tres niveles de detalle:

- Clases y propiedades de punto de partida: proporcionan la base de la ontología PROV y se utilizan para crear descripciones simples de procedencias en donde únicamente se describen las entidades, los agentes y las actividades.
- Clases y propiedades expandidas: proporcionan términos adicionales para describir de un modo más detallado aspectos relacionados con la procedencia tales como personas, agentes de software, colecciones, descripciones de procedencia y localizaciones.
- Clases y propiedades cualificadas: Describen aspectos muy detallados relacionados con la

influencia entre entidades, agentes y actividades, revisiones, delegaciones, asociaciones entre agentes, entidades derivadas, etc.

En ocasiones PROV-O resulta muy compleja de utilizar. La ontología PAV ofrece un enfoque más sencillo y al mismo tiempo complementario. PAV especializa algunos términos de PROV-O y *Dublin Core* para describir la autoría, la curación y la creación digital de recursos en línea. Aunque *PAV* entra dentro de lo que se denomina "ontología ligera" proporciona los términos para distinguir entre los diferentes roles que desempeñan los agentes en relación a un recurso, tales como la creación, importación, derivación, curación, recuperación o versionado entre otros (**Ciccarese**; **Soiland-Reyes**, 2015).

"Los usuarios precisan contar con mecanismos para citar conjuntos de datos o conocer las herramientas de acceso disponibles"

Calidad

La calidad de un conjunto de datos se mide en función de determinadas propiedades de la totalidad del conjunto de datos y de sus distribuciones. *DQV* (*Data Quality Vocabulary*) es una extensión del vocabulario *DCAT* para representar aspectos tales como la frecuencia de actualización, si se aceptan cambios sugeridos por los usuarios, si existe un compromiso por parte de los editores para mantener y preservar el conjunto de datos, etc. (**Albertoni**; **Isaac**, 2016). Este vocabulario puede ayudar a los programadores de aplicaciones a confiar en los datos que utilizan las mismas. Para expresar las propiedades de calidad de un conjunto de datos o una distribución *DQV* define cinco tipos de información (con sus correspondientes clases y propiedades):

- retroalimentación, puntuaciones, calificaciones y certificaciones de calidad;
- conformidad con respecto a los estándares utilizados;
- políticas o acuerdos regidos por criterios referidos a la calidad de los datos;
- métricas de calidad cualitativas o cuantitativas;
- entidades involucradas en la procedencia del conjunto de datos o distribución.

DQV define las medidas de calidad a partir de una serie de dimensiones de calidad y métricas de calidad:

- las dimensiones hacen referencia a determinadas características;
- las métricas definen procedimientos para medir las dimensiones de calidad.

Uno de los aspectos más interesantes de este vocabulario es el mecanismo de derivación de las anotaciones de calidad. DQV parte del principio de que una información relativa a la calidad puede derivarse a partir de otra. Por este motivo DQV permite derivar anotaciones sobre la calidad de un conjunto de datos o distribución a partir de estándares o medidas de calidad ya existentes o definidas con anterioridad.

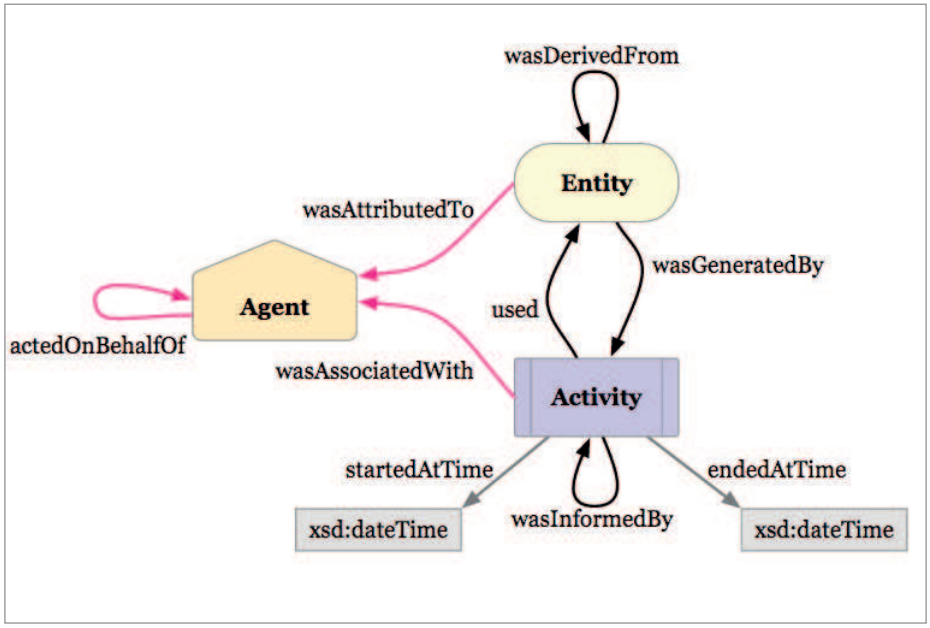

Figura 2. Elementos principales de la ontología PROV
https://www.w3.org/TR/2013/REC-prov-o-20130430/diagrams/starting-points.svg

Uso, citación y retroalimentación

Los datos publicados en la Web pueden ser utilizados por los consumidores de múltiples formas. Sobre este punto resultaría de gran interés que los editores pudieran disponer información sobre la experiencia de uso de dichos datos. Por su parte los usuarios precisan contar con mecanismos para citar conjuntos de datos o conocer las herramientas de acceso disponibles. DUV (*Data Usage Vocabulary*) se ha definido con esta finalidad,

proporcionando un modo para que los usuarios puedan citar datos, metadatos descriptivos sobre cómo acceder y utilizar conjuntos de datos y distribuciones, así como una forma para representar los comentarios y experiencias de los consumidores de datos (**Farias-Lóscio**; **Stephan**; **Purohit**, 2016).

"Existe un área de gran interés en la investigación sobre modelos de arquitectura de aplicaciones para la publicación y consulta de datos en la Web"

DUV tiene una estructura modular, y ofrece tres submodelos especializados: citación, uso y retroalimentación:

- Submodelo de citación: los conjuntos de datos y las distribuciones son formas de medios electrónicos que pueden ser citados por los consumidores de datos utilizando criterios bibliográficos básicos de referencia proporcionados por los editores de datos. Por otro lado, también pueden ser anotados por los editores de datos con referencias bibliográficas que proporcionan información adicional a los consumidores de datos e incluso con referencias proporcionadas por los consumidores de datos.
- Submodelo de uso: sería de gran utilidad realizar un seguimiento del uso que realizan de los datos los propios consumidores. Esto podría ayudar a mejorar la utilidad de los conjuntos de datos al describir cómo pueden ser utilizados y se incorporaría como metadatos descriptivos proporcionados por el editor de datos a la comunidad de consumidores.
- Submodelo de retroalimentación: proporciona una forma para compartir información de retroalimentación y evaluación de los datos. La retroalimentación del usuario es importante para solucionar problemas relacionados con la calidad de los datos publicados. Diferentes usuarios pueden tener experiencias diferentes con el mismo conjunto de datos por lo que *DQV* permite representar el contexto en el que se utilizaron los datos y el perfil del usuario que lo utiliza. La retroalimentación también permite que los consumidores comuniquen correcciones o sugerencias al editor.

Conclusiones

La publicación de datos en la Web aplicando criterios de interoperabilidad es algo que ya se da por supuesto. Puede afirmarse sin lugar a dudas que se ha superado la fase de desarrollo de tecnologías y la publicación experimental de conjuntos de datos. La aplicación de *linked open data* es algo que se da por sabido y la definición de conexiones, mapeados y enlaces RDF entre conjuntos de datos es una práctica bastante común. Las soluciones de software en forma de *frameworks* de desarrollo y almacenamiento de datos RDF ya es un campo bastante trillado.

No obstante, todavía queda un terreno muy fecundo en el campo de la investigación sobre modelos de arquitectura de aplicaciones para la publicación y consulta de datos en la Web. La aplicación de buenas prácticas y la creación de vocabularios para la representación de metadatos que describan el contenido, la estructura, procedencia, calidad y uso de conjuntos de datos permitirá definir con mayor precisión las características que deben incorporar las plataformas de publicación de datos en la Web. A este respecto, en un futuro inmediato asistiremos a una intensa actividad de elaboración, reelaboración, simplificación y despliegue efectivo de vocabularios como DCAT, PROV-O, DQV o DUV. La calidad de los conjuntos de datos ya no se medirá únicamente a partir de la calidad de los datos que contiene, sino también mediante los metadatos que permiten

Incluso es posible pensar en un horizonte más amplio, ya que muy posiblemente en un futuro cercano la publicación y descripción de datos en la Web según un conjunto de buenas prácticas es algo que se dará por supuesto. En ese momento la clave será el uso que se haga de los mismos mediante APIs y aplicaciones desarrolladas por terceros.

Notas

1. Se utiliza la expresión "vocabulario" para referirse a un conjunto de clases y propiedades definidos de forma conjunta. Resulta especialmente complejo definir de forma excluyente y delimitada los conceptos a los que se refieren expresiones como ontologías, esquemas de metadatos, términos de vocabularios, etc. Sobre este punto, tal y como indica el *W3C* no hay una división clara y el nivel de complejidad tal vez sea el factor determinante para utilizar una expresión u otra: *https://www.w3.org/standards/semanticweb/ontology*

2. *LOV*, *Linked Open Vocabularies*, es una iniciativa que permite registrar vocabularios de metadatos y ontologías. Más información: *https://lov.okfn.org*

3. URI e IRI: Una referencia IRI es una generalización de las referencias URI que permite una gama más amplia de caracteres *Unicode*. Esto permite la codificación de las referencias IRI mediante caracteres UTF-8 que en un principio puede que no estén permitidos en las referencias URI. Más información en: *https://www.w3.org/TR/rdf11-concepts/#dfn-iri*

4. Generalmente el servidor está configurado para que, en el caso de que se soliciten recursos asociados a una IRI que ya no está disponible (pero que existió en el pasado), devuelva el código de estado "410 Gone" (para

informar simplemente que dicha IRI ya no está disponible) o "303 See Other" para ofrecer una localización con información adicional sobre dicho recurso en el caso de que su IRI haya cambiado o se haya archivado.

5. Más información en:
http://dublincore.org/documents/2010/10/11/dcmi-terms

6. Más información en:
http://xmlns.com/foaf/spec/

Referencias

Albertoni, Riccardo; **Isaac, Antoine** (2016). *Data on the Web best practices: Data quality vocabulary. W3C Working Group Note 30 August 2016.*
https://www.w3.org/TR/2016/NOTE-vocab-dqv-20160830

Alexander, Keith; **Cyganiak, Richard**; **Hausenblas, Michael**; **Zhao, Jun** (2011). *Describing linked datasets with the VoID vocabulary. W3C Interest Group Note 03 March 2011.*
http://www.w3.org/TR/2011/NOTE-void-20110303

Bizer, Christian (2007). *Quality-driven information filtering in the context of Web-based information systems.* Tesis Doctoral. *Freie Universitat Berlin.*
http://www.diss.fu-berlin.de/diss/servlets/MCRFileNodeServlet/FUDISS_derivate_000000002736/

Carothers, Gavin (2014). *RDF 1.1 N-Quads: A line-based syntax for RDF datasets. W3C Recommendation 25 February 2014.*
https://www.w3.org/TR/2014/REC-n-quads-20140225

Ciccarese, Paolo; **Soiland-Reyes, Stian** (2015). *PAV - Provenance, authoring and versioning.*
http://pav-ontology.github.io/pav
http://purl.org/pav/2.3

Cyganiak, Richard; **Zhao, Juan**; **Alexander, Keith**; **Hausenblas, Michael** (2011). "Vocabulary of Interlinked Datasets (VoID)". *Deri Vocabularies.*
http://vocab.deri.ie/void

Farias-Lóscio, Bernadette; **Burle, Caroline**; **Calegari, Newton** (2016). *Data on the Web best practices. W3C Candidate Recommendation 30 August 2016.*
https://www.w3.org/TR/2016/CR-dwbp-20160830

Farias-Lóscio, Bernadette; **Stephan, Eric G.**; **Purohit, Sumit** (2016). *Data on the Web best practices: Dataset usage vocabulary. W3C Working Group Note 30 August 2016.*
https://www.w3.org/TR/2016/NOTE-vocab-duv-20160830

Lebo, Timothy; **Sahoo, Satya** (2013). *PROV-O: The PROV ontology. W3C Recommendation 30 April 2013.*
http://www.w3.org/TR/2013/REC-prov-o-20130430/

Maali, Fadi; **Erickson, John** (2014). *Data catalog vocabulary (DCAT). W3C Recommendation 16 January 2014.*
http://www.w3.org/TR/2014/REC-vocab-dcat-20140116

Juan-Antonio Pastor-Sánchez
Universidad de Murcia
pastor@um.es

https://scholar.google.com/citations?user=zv5d900AAAAJ